Effects of changes in data collection mode on data quality in administrative data: the case of participation in programmes offered by the German employment agency

Seysen, Christian

Empfohlene Zitierung / Suggested Citation:
Seysen, C. (2009). Effects of changes in data collection mode on data quality in administrative data: the case of participation in programmes offered by the German employment agency. *Historical Social Research*, *34*(3), 191-203. https://doi.org/10.12759/hsr.34.2009.3.191-203

Effects of Changes in Data Collection Mode on Data Quality in Administrative Data. The Case of Participation in Programmes Offered by the German Employment Agency

Christian Seysen *

Abstract: *»Der Einfluss des Datenerhebungsmodus auf die Datenqualität von Verwaltungsdaten. Eine Methodenstudie am Beispiel deutscher Arbeitsmarktdaten«*. Until administrative data are available as research datasets, they are passed through many organizational units and stored in different formats. The transformations of collected data to a data warehouse and further the integration of data from several operational sources to an integrated dataset for research projects include various mappings of identifiers and variables. A particular challenge arises, whenever one of the intermediate products changes. The resulting difficulties are not only technical in nature, but may well lie in aspects of the theoretical interpretation in a particular research context. For long-term research projects, it is essential to ensure comparability between several versions of this dataset. So the main task resulting from changes in the data sources is to ensure that observations of a former version of a research dataset can be identified after these changes. A case study of the Integrated Employment Biographies (IEB) is presented as an example of these problems. In a first step reasons for changes in the data sources and the methodological problems of transformation between several versions of a research dataset are highlighted. In a second step some tests of variables fundamental for research analysis and stratification are presented.

Keywords: Longitudinal Analysis, Process-Generated Data, Social Bookkeeping Data, Public Administrational Data, Data Management, Data Collection Mode Record Linkage, Data Fusion, Labour Market Data.

1. Introduction

The increasing usage of process-generated data as a source of research in social sciences since the beginning has been connected with the discussion of methodological problems (Müller: 1977). The expansion of the data bases and the resulting possibilities of research projects have been discussed in sociological discourse as well as problems of data quality and methods of improvement (Karstedt-Henke 1984). One of the results of the discussion was that particu-

* Address all communications to: Christian Seysen, IT and Information Management, Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg, Germany; e-mail: christian.seysen@iab.de

larly the evaluation of information, the weighting of sources and the problems of aggregation and coordination of datasets require detailed knowledge of the process of data generation, which in many cases is only available for certain persons inside the data processing organisation (Karstedt-Henke 1984: 160). To ensure the reproduction of results based on the analysis of process-generated data, social scientists have developed a doctrine of errors ("*Fehlerlehre*"), which covers the whole process of data generation from data collection to data transformation and data storage. This doctrine of errors is based on a theory of the organisation as information processing system with its internal processes of information management and its relationships to the relevant environment of the organisation (Karstedt-Henke 1984: 161).

Using this model, the following discussion will treat effects of changes in data collection mode on data quality in administrative data. Questions of the quality in the case of process-generated data often arise after changes in the organisational structure or the technical infrastructure of the underlying processes. Due to a missing elaborated theory and appropriate models of the affected processes, in many cases occasional and limited activities of improving data quality have to be performed to ensure the constancy of analyses.

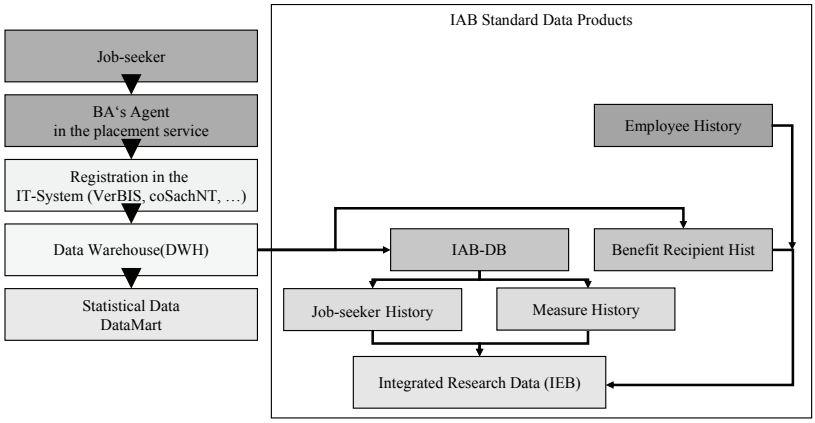
Some of the problems discussed in this paper can only be solved for specific data, as the researcher has to consider the specific character of the data. I illustrate this point and develop a suggestion for solving these problems by means of a case study. In this case study, I will focus on some aspects of data quality that are specific to a certain type of research data sets basing on German labour market data: .After an overview of an exemplary process of data generation and the organisational context (2), some types of changes in the process of data generation and their effects on research data sets will shortly be outlined (3). The leading point is to specify criteria of data quality regarding process-generated data and to exemplify benchmarks. The case study will exemplify structural changes in the process of data generation, (4.) using German labour market data as an example. Consequences are treated regarding methodological problems of the stability of data (5.) and basal measures of quality assurance concerning this matter (6).

2. The Process of Data Generation in German Institute for Employment Research

Administrative data and data quality in the following will be regarded from a certain point of view, which is determined by the institutional and organisational context of the Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung. Die Forschungseinrichtung der Bundesagentur für Arbeit IAB). The IAB as a Research Institute of the Federal Employment Agency in Germany uses data which are collected in administra-

tive processes of the local agencies. Figure 1 gives an overview of the process of data generation and the structure of the data products at the IAB.

Figure 1: Data Collection Process in the BA and Generation of Dataproducts at thte IAB



These administrative processes are initiated by clients who come to the Agencies to register as job-seekers. They apply for benefit, look for employment services and possibly they participate in measures of employment and training schemes. The data on these activities are collected in computer systems in the employment agencies and forwarded to central databases in a Data Warehouse (DWH). There, the data are transformed to a historical database of business data. The DWH is the point of reference for statistical reporting of the Federal Employment Agency (Bundesagentur für Arbeit BA) as well as for IAB Standard Data Products, which provide data for research programs. However, this description of the data production process is an ideal, since not all of the administrative data for research programs are available in the DWH yet.

The following methodological discussion and the case study focus on research data sets of the IAB which have a particular place in this context. Research projects of the IAB or external research institutions do not access the data directly from the DWH as the central databases of the BA. Instead, by law it is the task of the IAB, and particularly of the department IT and Information Management (IT und Informationsmanagement ITM) to develop and store research datasets on the base of the data, which are generated by BA (Köhler/Thomsen in this special issue). To meet these demands in an effective and efficient way, the IAB developed some standard data products which are periodically updated: In a first step, the IAB creates historical research data sets which each are based on a specific administrative procedure like the Benefit Recipient History (Leistungsempfänger-Historik LeH). In a further step, data

from diverse administrative procedures are brought together in the Integrated Employment Biographies (Integrierte Erwerbsbiografien IEB).

Looking on the latter type of data, the following methodological discussion can be specified: The IAB receives data from the DWH, which originally are collected by administrative processes and which were designed for these operational tasks. The conception of the standard data products of the IAB is to provide historical research data sets which are updated regularly. These historical data sets contain information about administrative processes in the past. So in view of constancy, it would be ideal, if a new version of a dataset would contain exactly the same information for the same periods in the past and only differ from a former version in additional time periods. On the other hand, the data should be modified, if new information on data which were collected in the past arises. E. g. new information about the type of a training scheme in the past may be collected in the local agency, that a client participated in a training scheme, not in language course. This newly collected information produces a new record in the DWH and the corresponding record in the IAB-Database should be updated. Another example: changes in technical keys, which represent the differentiation of training schemes in the collecting computer system should not produce another information in the research dataset about a participation in training schemes in the past.

Although this might seem self-evident and as a simple problem, if one looks at the processes of building research datasets, it becomes clear, that this is actually not as easy at all: Since the datasets base on the process of data generation in the BA, they are potentially affected by each modification of the process. This may be changes, which bring additional information, but the consequence also can be the loss of information. From this point of view some problems of data quality of these datasets will have to be discussed in a specific way.

3. Changes in Data Collection Mode and Criteria of Data Quality

The problem of data quality is an important point of the discussion of process-generated data (Karstedt-Henke: 1984). Regarding the doctrine of errors mentioned above a main point is the relationship between changes in the context of data generation and information management of an organisation (Karstedt-Henke: 162). It is not the aim of this article to discuss this relationship in detail, but to focus on consequences of changes in the data collection mode for research datasets.

Most of the changes in data collection mode are connected with the implementation of a new technical data collection system, mostly initiated by a modification of a law and a new conception of the administrative processes, the

data are related on. So the reasons of changes are functional or technical, in most cases it is both.

These changes in data collection mode and the effects for the whole process of data generation can have different extensions. They can affect single values in the data collection, e. g. when a collected attribute is more differentiated and additional values are defined. Changes of larger extent are connected with the implementation of new software for data collection or a modification in the structure of data management.

At the level of the resulting datasets, effects concretely can be missing or undefined values, errors in calculated values, incomparability of variables or incomparability of identifiers. To judge these effects regarding to the data quality of the resulting datasets, specific criteria of data quality have to be established.

The methodological discussion of *quality criteria* of data in many cases refer to the classical test theory as a statistical theory of errors (Schnell et al. 2005: 149). To identify and estimate measurement errorss, this theory bases on axioms such as the correlation between a measurement and its object or between a real measurement and a 'true value'. In view of the classical test theory, fundamental quality criteria of measurements were established, like objectivity, reliability, validity, comparability and efficiency (Schnell et al. 2005: 151). The basal criterion 'objectivity' is related to the independency of a measurement from the person who takes it. As most important criteria reliability and validity are often mentioned, which are related in the research process: The reliability as temporally stability is a precondition of the validity, since a measurement cannot actual measure what it is intended to measure, if there is no stability in the iteration of the measurements.

Regarding to standard datasets for research and the constancy of data over several actualisations of a dataset, the following discussion will focus on the aspect of reliability. *Reliability* in this case means the stability of standardized research data products, which are updated periodically.[1] This question would lead to conceptional problems of standard data products, which on the one hand enable a practical data management and on the other hand address most of the requirements of research projects.

The classical test theory defines criteria for reliability and the relationship to criteria of validity as statistical values. Two main assumptions are that measurements errors are random errors and that the values of measurements are

[1] I will not discuss objectivity, since this seems to be not the crucial point in the case of changes in the process of data collection und transformation: Once implemented, changes of a process establish a new structure of data transformation and as such the transformations should be independent from a collecting person. On the other side, I will not discuss validity, since the decision which data are valid depends on the researcher's research question, concepts and theory, relating to which the (Kruppe in this special issue).

randomly distributed around a 'true value'. The additional assumption that errors are not correlated with 'true values' and errors of several actual measurements are not correlated allows a statistical estimation of reliability and validity (Diekmann 2007: 261-270).

Regarding standard data products based on process-generated data, the question of reliability must be specified. On the one hand, the temporal stability of measurements based on an implemented process is not a problem as long as the process of data generation is not changed, since the structure of data transformation remains constant. On the other hand, errors caused by a data transformation in the most cases are not randomly distributed, but cause systematic biases.

The question of reliability particularly arises after changes of the process of data generation. The crucial point of reliability in this case is the question of constancy and comparability between different datasets which base on the same observations, but which are results of different and changed processes of data transformation. E. g. changes in the collecting computer system and the transformation processes may modify keys for the differentiation of training schemes. But the researcher in an evaluation project has to identify the participations in a certain measure in the datasets before and after these changes.

4. Structural Changes in Data Transformation

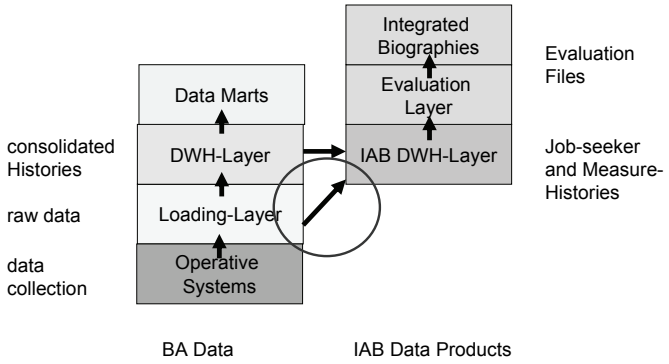
Regarding to the introduced organisational structure and the correspondent architecture of data management, a case study can illustrate the problems of data transformation which require an improvement of data quality. The following chart shows the way data are transformed from operative systems to the central data warehouse of the BA and further to the IAB standard data products.

The chart presents the typical steps that have to be taken in order to maintain the transformation of administration data (in this case: BA data) to research data (in this case: IAB data) after changes in the data collection mode. After changes in the collecting computer systems, the new data structure has to be provided in the data warehouse-layer for further usage. In the presented case the data were not available in the central DWH-Layer of the BA in time to deliver data for a research project. The circled arrow illustrates the requierement of temporal solutions in the IAB to ensure the delivery of research data in such situations.

After changes in the collecting computer systems for the administration of employment measures the IAB had to build up an own data warehouse-layer as database for the IAB standard products. This database had to base on raw data of the loading layer, where the data from the local agencies are stored, and had to implement the transformations to historicised data in the DWH-Layer. It is obvious that changes to such an extent in one step of the transformation process

affect the following steps. So it is important to measure effects of these changes in the data generation on the resulting data products. To ensure data quality of standard products at the end of the chain of transformations it must be generated benchmarks regarding to the criteria of data quality, which were mentioned above.

Figure 2: Transformation from Administrational Files to Research Files



The case study I will present below illustrates changes in the process of data generation and minimal criteria of assurance of constancy. In view of reliability over different versions of a dataset, three main requirements of analyses were proposed:

1) *Identification of Single Cases in Single Data Sets:* Basal for the identification of an observation is the stability of identifiers of persons and their participations in measures (i.e. employment and training schemes)
2) *Reidentification of Single Cases in Across Data Sets:* Beside the identifiers even the stability of the date specifications of measures and participation in measures is an important requirement reidentification after technical changes.
3) *Constancy of Measures:* The constancy of the types of measures has to be established.

All these attributes are fundamental for analyses. After the changes in the data warehouse-layer and evaluation layer as source of the Integrated Employment Biographies (IEB), the standardised assurance of data quality was extended by special tests. The aim was to get benchmarks for continuity and discontinuity of data regarding participations in measures of active labour market policy in the IEB: The changes of the data warehouse-layer and of the rules of data transformation to the evaluation layer affected the identification of observations itself. Due to the redesign of the data transformations it was to expect a difference of the total numbers of participants and participations in

measures, as well as modifications of the time period of a participation. So the data quality assurance had to assess the three points stated above by the following procedures:

1) *Reidentification of Single Cases in Single Data Sets:* A first measure of assessing reliability ost to test the total numbers of participants in the IEB before and after the changes, as well as the numbers of reidentifiable, missing or new participants for the same time periods in both versions of the dataset.
2) *Reidentification of Single Cases in Across Data Sets:* Quality Assurance also had to test total numbers and reidentifiability of participations.
3) *Constancy of Measures:* Since the type of a measure of active labour market policy is a main variable of analyses, the distributions regarding this variable in both versions of the IEB were compared.

This test scenario was a minimal approach to get benchmarks of the criterion reliability or constancy respectively. [2]

5. Assessing Reliability: An Example Using IAB Data

I now will present a detail from data transformation which particularly illustrates the problem of reidentification of observations and the effect on constancy of basic variables in the case of process-generated labour market data.

Table 1: Example Cases

Participations in Measures (IAB-DB MTG)							
ID	**Part-ID**	**Client-ID**	**Start**	**End**	**Typ**	**Status**	**Collection**
10001	30001	50001	01.06.2003	31.05.2004	lang. course	entry	15.05.2003
10002	30001	50001	10.06.2003	31.05.2004	training	entry	01.06.2003
10003	30001	50001	10.06.2003	31.05.2004	training	stock	15.06.2003
10004	30001	50001	10.06.2003	31.05.2005	training	stock	01.10.2003
10005	30001	50001	10.06.2003	31.03.2004	training	stock	01.03.2004
10006	30001	50001	10.06.2003	31.03.2004	training	leaving	01.04.2004
10007	30001	50001	10.06.2003	31.03.2004	training	leaving	25.04.2004

Participations in Measures (MTG Evaluation Layer)							
id	part_id	client_id	b_start	b_end	e_start	e_end	typ
90001	30001	50001	10.06.2003	31.05.2004	10.06.2003	31.03.2004	training

[2] The analysis of Engelhard et al. (2008) contains a detailed description of the data collection process and the transformations to the IAB-Databases before and after the change of the data collection mode. The main point of view is an analysis of continuity and discontinuity of the data respectively, focussed on the measure types of the ESF-BA Program. The main result in the point of continuity is similar to the quality assurance report of the case study in this article.

Table 1 shows a fictitious detail from a dataset in the IAB DWH-Layer containing Participations in Measures of Active Labour Policy (in this case: training schemes) - before the redesign. All records belong to one client (Client-ID) and one participation (Part-ID). The records are collected within a time period of nearly one year and contain key information about the time period of the participation in measure and the type of the measure. Further key variables are the status of a record as collected at the beginning of the participation in measure, during the participation or at the end ('entry', 'stock' or 'leaving') and the collection date. The chart only shows a selection of basic variables of analyses, which play a role in the discussed problem – the actual data base contains a lot more variables.

5.1 Identification of Single Cases in Single Data Sets

Regarding the model introduced above, the next step of data generation is to build up the evaluation layer from these data. The transformation from the IAB DWH-Layer to the Evaluation Layer has to select one record for every participation of a person in employment measures.

Methodologically, this is the important step of identifying an observation in the historical data of a data base. Historized data in a data warehouse represent information including the time of their validity (in this case: the time period of a participation in an employment measure) and the time of their collection (in this case: the date of data collection in the local agency). The information about a person's participation in employment measures can vary at different times of data collection. Every time when an information about a participation in an employment measure is collected in the local agency a record is stored. For the most participatiants there are a lot of records with different informations at different times. So for analyses, it is important, in compliance with which administrational rule the information is selected. After changes in data collection mode, it is an essential question, whether the variables which are relevant for identification of an observation and the selection of information, still are available.

In the case of the IAB-database of participation in measures before the changes of the system, the identification of a participation for the Evaluation Layer refered to the collection status of a record as stored in the 'status' key. The information for the Beginning and the end of a participation in a measure was taken from the last collected record at the beginning (status: 'entry') and the last collected record at the end (status: 'leaving'). A problem for this type of identification arises, if changes of the system affect these keys of identification. This was the problem in the presented case:

5.2 Reidentification of Single Cases Across Data Sets

After changing the data collection system and the rules of loading data into the central DWH of the BA, the status key for marking the context of collection of data at the beginning, during the participation or at the end of a measure was not longer available. For the redesign of the IAB-database, the rules of selection of information regarding a participation in the case of multiple records for a participation could not relate to a key of the status of collection, but had to establish rules of selection only regarding collection time and start/end date of a participation in measure. Under this rules the information regarding the beginning of a participation was taken from the first collected record. Table 1 shows the case, that the different rules select different information about the beginning of the participation.

These modification of the transformation rules had different results not only for the date specification of the time period, but also for the type of the measure. Table 1 shows the typical case, that an information collected at time period 1 is corrected at time period 2. The type of the measure is updated from 'language course' in the first record to 'training' in the second record. Different rules of identification in this case can result in different information about the type of a measure.

As mentioned above the presented case of changes in the data transformation system caused a test programm for the resulting datasets to deliver benchmarks regarding a minimum of quality criteria. In view of the continuity of the basal identifiers in a first step the total number of persons and participations in the resulting dataset 'Integrated Employment Biographies' (IEB) were counted, once in the version before the changes of data generation (IEB v3), once after the changes (IEB v4). Even for the same time period of data collection in the new system the counted participations were higher. The following table contains the numbers of participations in measures of actiive labour policy differentiated (in the rows) by the main categories of the measures. These categorgies correspond to the underlying programs of labour policy and to the procedures of the collecting computer systems.

Table 2: Differences Between Source Files

Source	IEB v3	IEB v4	Difference	%
ABM	2.618.696	3.106.860	+488.164	18,64%
FbW	4.353.550	5.564.370	+1.210.820	27,81%
FF	1.244.165	1.282.768	+38.603	3,10%
ESF-BA	197.686	289.057	+91.371	46,22%

An analysis showed mainly two reasons: first a higher degree of data completeness due to a newer loading from the sources and second a less restrictive import of data into the evaluation layer. The second point relates to the discussed changes in the data transformation. At the beginning of the data collection period there are a lot of participations without a record with the status 'entry'. In the former version of the evaluation layer these participations were not imported. Since this status key does not exist in the new system, it is not a criterion for exclusion any longer. So the modification of the rules expands the amount of the database, which can be judged as quality intensification.

5.3 Constancy of Observations

A great challenge of the quality assurance activities was the test of the reidentifiability of participations in the dataset IEB before the changes (V3) in the dataset afterwards (V4). A connected problem was the analysis of the accordance of the participation time periods. These analyses are important and difficult as well, particularly if the reidentification of participations after a system change cannot base on technical keys and the collected functional keys not in any case are entirely filled.

In a first step, participations for the comparison of the datasets were identified by the basal identifiers of the collected data: the client-number for the participating person, , the measure-number for the concrete program the person is participating in and and the participation-number as identifier of the concrete participation of the person . In four out of five groups of types of measures, this way to identify a participation in measures in the data in the initial dataset (V3) affords an acceptable result. In the case of training schemes one of the basal identifiers, the measure-number, is not filled by the administrative procedure. So for this type of measures, the degree of unambigously identifiable participations without relation to an additional criterion is much lower then in the case of the others (see table 3).

Table 3: Identification of participations in measures across datasets by person, measure-number and participation-number

Type	V3 total	V3 unambiguous		V3 in v4		V3 not in V4	
	N	N	%	N	%	N	%
ABM	3.160.984	3.160.904	100,00%	3.116.403	98,59%	44.501	1,41%
FbW/DSL	2.163.134	2.161.604	99,93%	2.153.546	99,63%	8.058	0,37%
TM	**3.720.632**	**2.534.578**	**68,12%**	**2.534.555**	**100,00%**	**23**	**0,00%**
FF	1.712.698	1.712.696	100,00%	1.635.840	95,51%	76.856	4,49%
ESF-BA	268.762	268.285	99,82%	267.979	99,89%	306	0,11%

To increase the number of unambiguously identifiable participations in training schemes, the identification can refer to the time period of the participation as an additional criterion. However it is not possible to analyse the quality of the time information itself in this way. To enable this analysis, the observations have to be identified before, without relating on the time specification, even if a part of participations of this type cannot be analyzed in this way.

6. Results

The discussion of data quality of administrative data presented a small detail of data transformations and relating problems after changes in the procedures. A systematic quality assurance of research data basing on a complex architecture of data transformations is a great task and often it is more a regulative idea than a fulfilled standard. The aim of the presented case study was to give an idea of basal requirements which have to be ensured for a conception of standard data products like from the IAB provided.

The study focussed on problems after changes in the underlying processes of data generation and transformation. Basing on the discussed basic criteria of data quality, particularly the criterion of reliability, activities of data quality assurance were established. They ensured that the deviations in the research data set after the redesign of the IAB-database were in a tolerable range.

The discussion shows, that data quality is related to technical aspects of data transformation as well as to underlying administrative processes of data collection, the rules of data transformation and the concepts of evaluation. The experience of the presented activities of quality assurance confronted with a specific dilemma regarding the design of standard data products: Data transformations of process-generated data often are related to technical or functional keys which change with a new system. In some cases the reliability of research data based on process-generated data after changes of the data collection mode can only be ensured by complex technical transformations and mappings. But this can make it difficult to judge the validity. The other way aground: the redesign of standard data products after changes in the underlying processes sometimes allow better transparency of validity of the data, but may result in less constancy in the research datasets

References

Diekmann, Andreas (2007): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, Reinbeck: Rowolt

Engelhardt, Astrid / Oberschachtsiek, Dirk / Scioch, Patrycja (2008): Datengenese zweier Datenkonzepte. MTG (Maßnahme-Teilnahme-Grunddatei) und ISAAK (Instrumente Aktiver Arbeitsmarktpolitik). Eine Betrachtung ausgewählter Fälle

am Beispiel der Förderung im Rahmen des ESF-BA-Programms. FDZ Methodenreport, 08/2008.
Karstedt-Henke, Susanne (1984): Die Entwicklung von Prüfverfahren bei der Verwendung prozeß-produzierter Daten. In: Bick, Wolfgang / Mann, Reinhard / Müller, Paul J. (1984): Sozialforschung und Verwaltungsdaten. Historisch-Sozialwissenschaftliche Forschungen (HSF). Stuttgart: Klett-Cotta.
Kruppe, Thomas und Martina Oertel (2003): Von Verwaltungsdaten zu Forschungsdaten – Die Individualdaten für die Evaluation des ESF-BA-Programms 2000 bis 2006. IAB Werkstattbericht Nr. 10.
Müller, Paul J. (Ed.) (1977): Die Analyse prozeß-produzierter Daten. Historisch-Sozialwissenschaftliche Forschungen (HSF). Stuttgart: Klett-Cotta.
Scheuch, Erwin K.: Die wechselnde Datenbasis der Soziologie. Zur Interaktion zwischen Theorie und Empirie. In: Müller, Paul J. (Ed.) (1977): Die Analyse prozeß-produzierter Daten. Historisch-Sozialwissenschaftliche Forschungen (HSF). Stuttgart: Klett-Cotta. 5-41.
Schnell, Rainer / Hill, Paul R. / Esser, Elke (2005): Methoden der empirischen Sozialforschung. München/Wien: Oldenbourg Verlag.